互联网内容审核

职业技能等级标准

中国电子学会

2020 年 5 月 发布

目录

前　言

本标准按照 GB/T 1.1-2009 给出的规则起草。

本标准根据《国家职业教育改革实施方案》的要求，以及《中华人民共和国劳动法》的有关规定，为进一步完善职业教育标准体系，为职业教育、职业培训和职业技能鉴定提供科学、规范的依据，标准编写组在广泛调查研究基础上，并征求了有关单位和专家的意见，经反复讨论、修改和完善，制定了《互联网内容审核职业技能等级标准》（以下简称本《标准》）。

本标准主要起草单位：中国电子学会、北京中工信推信息技术中心

本标准参与编写单位：中国信息安全测评中心、北京百度网讯科技有限公司、华为技术有限公司、网易（杭州）网络有限公司、北京中科汇联科技股份有限公司、浙江齐聚科技有限公司、同盾科技有限公司、深圳联想懂的通信有限公司、北京网景盛世技术开发中心、中联国创（北京）科技发展有限公司等。

本标准主要起草人：李宝民、王宗杰、牛少章、王海平、徐建、姜园、陈晨、柳海峰、任望、祝卓、游世学、蒋韬、陈鑫璐、刘瑞霞、李文祥、张涛、王赓、徐利峰。

1. 范围

本标准规定了互联网内容审核职业技能等级对应的工作领域、工作任务及技能要求。

本标准适用于从事互联网内容审核工作人员职业技能等级的考核与评价，互联网内容审核从业人员的聘用、教育和职业培训可参照使用。

2. 规范性引用文件

下列文件对于本文件的应用是必不可少的。凡是注日期的引用文件，仅注日期的版本适用于本文件。凡是不注日期的引用文件，其最新版本适用于本文件。

GB/T 19001-2016 质量管理体系要求

GB/T 19000-2016 质量管理体系基础和术语

GB 17859 计算机信息系统安全保护等级划分准则

GB/T 22240 信息安全技术信息系统安全等级保护定级指南

GB/T 25069 信息安全技术术语

GB/T 31167—2014 信息安全技术云计算服务安全指南

ISO/IEC 17000：2004 合格评定—词汇和通用原则

ISO/IEC 17021-2.2：2009 管理体系审核认证机构的要求及管理体系第三方认证审核的要求—第 2 部分：管理体系审核第三方认证审核的要求

3. 术语和定义

国家、行业标准界定的以及下列术语和定义适用于本文件。

3.1 内容审核 Content Moderation

内容审核是指对互联网网站及各类平台、工具等承载的内容进行审查，并对部分内容进行监视、过滤和删除的行为。

3.2 智能审核管理系统 Intelligent Moderation Management System

智能审核管理系统是指采用人工智能技术的互联网内容审核管理系统。

3.3 自发性知觉经络反应 Autonomous sensory meridian response（ASMR）

自发性知觉经络反应是指人体通过视、听、触、嗅等感知上的刺激，在颅内、头皮、背部或身体其他部位产生的令人愉悦的独特刺激感，又名耳音、颅内高潮等。

3.4 接口 Interface

接口是系统与另一系统（或系统的某些部分）之间的公共边界，信息通过该公共边界传递。

3.5 检测引擎 Detection Engine

检测引擎是指根据一定的策略、运用特定的计算机程序从互联网上采集信息，在对信息进行组织和处理后，为用户提供信息检测服务，将内容审核的相关信息展示给用户的系统。

3.6 光学字符识别 Optical Character Recognition（OCR）

光学字符识别是指电子设备（例如扫描仪或数码相机）检查纸上打印的字符，通过检测暗、亮的模式确定其形状，然后用字符识别方法将形状翻译成计算机文字的过程。

3.7 内容分发网络 Content Delivery Network（CDN）

内容分发网络是构建在现有网络基础之上的智能虚拟网络，依靠部署在各地的边缘服务器，通过中心平台的负载均衡、内容分发、调度等功能模块，使用户就近获取所需内容，降低网络拥塞，提高用户访问响应速度和命中率。

3.8 统一资源定位符 Uniform Resource Locator（URL）

统一资源定位符是对可以从互联网上得到的资源的位置和访问方法

的一种简洁的表示，是互联网上标准资源的地址。

3.9 自然语言处理 Nature Language Processing (NLP)

自然语言处理是从破解成功人士的语言及思维模式入手，独创性地将他们的思维模式进行解码后，发现了人类思想、情绪和行为背后的规律，并将其归结为一套可复制可模仿的程序；也是关于人类行为和沟通程序的一套详细可行的模式。

3.10 深度学习 Deep Learning

深度学习是机器学习领域中一个新的研究方向，通过组合低层特征形成更加抽象的高层表示属性类别或特征，以发现数据的分布式特征表示。

3.11 图像识别 Image Recognition

图像识别技术是指对图像进行对象识别，以识别各种不同模式的目标和对像的技术。

3.12 用户生产内容 User Generated Content（UGC）

用户生产内容概念最早起源于互联网领域，即用户将自己原创的内容通过互联网平台进行展示或者提供给其他用户。

3.13 专业生产内容 Professional Generated Content（PGC）

专业生产内容是互联网术语，用来泛指内容个性化、视角多元化、传播民主化、社会关系虚拟化。

3.14 人脸识别 Face Recognition

人脸识别是基于人的脸部特征信息进行身份识别的一种生物识别技术。用摄像机或摄像头采集含有人脸的图像或视频流，并自动在图像中检测和跟踪人脸，进而对检测到的人脸进行脸部的一系列相关技术，通常也叫做人像识别、面部识别。

3.15 报文摘要算法 Message-Digest Algorithm 5

报文摘要算法（MD5）是一个将任意长度的数据字符串转化成短的固定长度的值的单向操作。

3.16 审核员 Auditor

审核员是指实施审核的人员。

3.17 能力 Competence

能力是经证实的能够应用知识和技能实现预期结果的特质。

3.18 职业技能 Professional Skills

职业技能指就业所需的技术和能力。

4. 面向工作岗位（群）

主要面向从事对互联网网站及各类平台、工具等承载的内容进行审查，并对部分内容进行监视、过滤和删除行为、智能审核平台及系统的开发、部署与管理、智能平台及审核系统的部署与配置、智能平台及审核系统的优化、审核计划的制定与实施、审核项目管理与运营策略、人才培训与指导等工作。

4.1 互联网内容审核（初级）

主要面向音视频、社交媒体、网络游戏、传媒、广告、电子商务、公共安全、公共事物、电子政务等领域互联网相关内容发布企事业单位从事文本检测、图片检测、视频检测、音频检测、人工审核、和智能审核系统的使用以确保其内容合规及安全。

4.2 互联网内容审核（中级）

主要面向音视频、社交媒体、网络游戏、传媒、广告、电子商务、公共安全、公共事物、电子政务等领域互联网相关内容发布企事业单位从事复杂功能逻辑和特殊场景需求的内容审核工作、智能审核系统的开发与管理、智能审核系统的部署与配置。

4.3 互联网内容审核（高级）

主要面向音视频、社交媒体、网络游戏、传媒、广告、电子商务、公共安全、公共事物、电子政务等领域互联网相关内容发布企事业单位从事复杂功能逻辑和特殊场景需求的内容审核运营策略和解决方案研发规划、智能审核系统的优化、审核计划的制定与实施。

5．面向院校专业领域

院校	专业大类	专业代码	专业类
中职	信息技术类	090100	计算机应用
		090500	计算机网络技术
		090700	网络安防系统安装与维护
		091500	通信技术
		091700	通信系统工程安装与维护
高职	电子信息大类	610201	计算机应用技术
		610202	计算机网络技术
		610211	信息安全与管理
本科	计算机类	080903	网络工程、软件工程
		50301	新闻学
		50304	传播学
		080904K	信息安全

6．职业技能等级标准

6.1 职业技能等级划分

互联网内容审核职业技能等级分为三个等级：初级、中级、高级，依次递进，高级别涵盖低级别要求。

“初级”定义为熟练使用各类审核工具/产品；

“中级”在初级的基础上需要掌握在内容审核规则配置平台上配置审核策略、在人机审核平台上合理分配审核角色，制定、管理审核项目的能力；

“高级”需要在中级的基础上掌握将监管规则落地到实际审核流程、规则的能力，能通过设计调整组织架构、制定审核计划以更好的适应审核规范；

6.2 职业技能等级标准描述

本部分描述的互联网内容审核职业技能各等级要求的内容，具体见下列表单：

表 1　互联网内容审核(初级)

工作领域	工作任务	职业技能
1.互联网内容安全	1.1 文本/视频/图片/音频检测	1.1.1 能掌握互联网审核相关法律法规，在规定时间范围内完成规定数量图片审核，准确率达到规定的数量； 1.1.2 能运用网络视听内容审核标准，支持直播、点播视频过检，保障视频内容安全； 1.1.3 能依据图像识别算法和大数据分析技能，精准过滤涉黄、推广、暴恐、政治类有害信息、虚假和其他个性化定义的违规图片； 1.1.4 能使用语音识别技术，结合反垃圾文本过滤规则体系，进行精准和高效的违规音频识别分析； 1.1.5 能掌握《微博客信息服务管理规定》，发现微博客服务使用者发布、传播谣言或不实信息，应当主动采取辟谣措施。

	1.2 短视频内容安全检测	1.2.1 能运用 LOGO 识别技术，进行短视频中 LOGO 检测，如台标、商品品牌等，防止侵权违规； 1.2.2 能利用视频指纹或者图像检索技术，反查互联网或其他竞品平台，是否存在侵权行为； 1.2.3 能掌握视频指纹、基于深度特征的视频检索等技术，进行快速资源排查，筛选违规内容，提高管控能力和效率。
	1.3 人工审核	1.3.1 能掌握机器判定疑似违规的内容，进行人工审核确认； 1.3.2 能具备实时接收审核要求、进行快速布控、快速审核的应急响应能力； 1.3.3 能正确了解互联网内容审核流程规范，使用审核工具，进行互联网内容审核； 1.3.4 对未违规/违法的网络舆情内容，若有必要，进行疏导建议。
2. 智能审核管理系统	2.1 智能审核管理系统运作	2.1.1 能正确了解关键词、黑白名单、过滤器和分类器的定义，进行系统操作； 2.1.2 能掌握智能审核管理系统的各项功能，进行基于内容特征识别（肤色、纹理）、贝叶斯过滤、相似度匹配和规则系统的内容审核； 2.1.3 能运用人工智能基础知识，进行大数据分析（用户行为、用户分类）、人机识别、机器学习（语义识别、图像识别）的相关智能审核操作。

	2.2 违规内容排查	2.2.1 能熟悉使用视频指纹、基于深度特征检索技术，对媒体库内容建立索引，快速进行资源排查，筛选违规内容，提高管控能力和效率； 2.2.2 能利用自定义文本、NLP 等知识，识别违规广告话术加入广告识别，依据广告法识别违规内容； 2.2.3 能掌握图像识别、视频分析、文本反垃圾等技术对发布内容做政治类有害信息、色情低俗、广告、引流、虚假等风险识别，防止违规内容主动传播、竞品打广告、引流。
	2.3 媒体资源库安全	2.3.1 能利用物体识别、场景识别知识，对视频进行打标签、分类，将整个媒资库结构化，以实现万物（物体、场景标签）识别； 2.3.2 能利用自定义文本、NLP 等技术对画面内容进行涉黄检查，除常规违规内容外，识别违规广告话术加入广告识别，依据广告法识别违规内容； 2.3.3 能使用 MD5/视频指纹技术，防止节目被篡改、编辑、插入、裁剪、翻拍；使用图片清晰度、美观度检测技术对视频质量进行审核。

3. 直播电商内容安全检测	3.1 直播电商内容安全检测	3.1.1 能运用《中华人民共和国广告法》基本合规要求对图片、文字、内容进行检测，如封面、公告栏、标题等； 3.1.2 能掌握《中华人民共和国电子商务法》相关法律法规，进行全面、真实、准确、及时地排查信息，筛选虚构信息，提高监管能力； 3.1.3 能掌握《中华人民共和国电子商务法》相关法律法规，筛选禁止销售的商品或服务；

表 2　互联网内容审核(中级)

工作领域	工作任务	职业技能
1.互联网内容审核	1.1 文本/视频/图片/音频检测	1.1.1 能掌握不同应用场景，进行文本评论/留言/跟帖中的涉黄、广告、灌水等敏感词或垃圾文本的实时高效过滤； 1.1.2 能依据不同场景，实时检测直播间内视频和 UGC/PGC 短视频的违规内容，进行截图过检和电视墙审核； 1.1.3 能利用不同应用场景，对用户头像、照片进行审核，过检违规、带有联系方式的图片； 1.1.4 能运用不同应用场景，对点播内音频内容传播类、IM 类语音、FM 电台类音频数据检测，保证音频数据合规性。

	1.2 保障直播内容安全	1.2.1 能结合人工智能审核技术检测直播画面中是否存在衣着暴露、裆部特写、臀部特写、性暗示动作、未成年人裸露、性爱玩具、营销推广、喝酒、吸烟吸毒、 赌博、误用国旗国徽、误用党旗党徽、出现恐怖组织旗帜标志、穿着军装、穿着宗教服饰等行为； 1.2.2 能运用语音识别和 NLP 技术，监测直播过程中主播是否存在涉黄、政治类有害信息等违规行为，降低直播低俗、政治类有害信息等内容风险； 1.2.3 能正确掌握人脸识别技术，进行直播主播和开播注册主播为同一人的身份确认，降低代播、未成年直播以及禁止黑名单人员直播风险。
	1.3 人工检测	1.3.1 能掌握审核基础知识与信息技术常识，对机器无法判定的疑似违规的内容进行人工审核确认； 1.3.2 能利用内容的价值与判断，对互联网信息的真实性、权威性、时效性作出及时判断； 1.3.3 能依据不同应用场景，对有害信息风险较高的以及有特殊审核标准的版块进行全量人工审核。 1.3.4 能灵活调整审核流程，对有害信息风险高和审核标准高及传播范围广的板块实行初审、复审和终审制。

2. 智能审核管理系统	2.1 智能审核管理系统的开发与管理	2.1.1 能依据各种场景的审核要求在智能审核管理平台中配置对应的审核维度、审核标签、审核松紧度，能对图像、文本、语音、人脸的自定义黑白名单进行增删改查的操作，能配置不同时期的审核策略，并实现到期自动切换； 2.1.2 能够结合不同业务场景的审核需求、法律法规，进行智能审核管理平台的功能升级与改善，包括但不限于：增加审核维度、增改审核标签、调整审核松紧度； 2.1.3 能结合人工智能审核的结论对审核内容进行验证，并设计合理的人工+机器审核工作流程，能在人机协同审核管理平台上配置审核任务，并下发给审核人员进行审核。
	2.2 智能审核管理系统的部署与配置	2.2.1 能掌握审核内容的需求与工作量，正确安装并布置智能审核系统； 2.2.2 能依据具体应用场景，正确安装具有复杂规则策略（包含关键词库、图片库、智能规则库、用户名单、高频规则库、IP 名单、文本特征库、设备名单、URL 名单和联系方式库等）匹配的智能审核系统； 2.2.3 能掌握审核进度，始终为所有任务组启用审核跟踪，配置和更改任何任务组的审核选项。

3. 互联网内容审核项目管理	3.1 项目管理	3.1.1 能掌握审核流程与进度安排，进行互联网内容审核人员审核工作的督导； 3.1.2 能依据具体应用场景，进行现场实际勘查、编写勘查报告和编制工作报告； 3.1.3 能利用现场管理和现场工作日志，组织质量讨论会，进行现场质量管理和编写项目总结报告。
	3.2 培训与指导	3.2.1 能掌握正确的互联网内容审核基础理论与实务，对初级人员进行培训； 3.2.2 能运用不同应用场景与案例，编写内容审核项目培训计划和实施技能测试； 3.2.3 能依据内容审核项目报告，指导初级人员进行错误原因分析和故障排除。

表 3　互联网内容审核(高级)

工作领域	工作任务	职业技能
1.互联网内容审核	1.1 文本/视频/图片/音频检测	1.1.1 能掌握互联网审核关键知识与技能，对审核工作提供技术支撑，其中包含但不限于：深度学习、NLP 技术（词向量模型、上下文语义识别）、内容变种智能识别与修正、行为识别、文本聚类等； 1.1.2 能依据智能截帧策略、用户行为分析、智能电视墙、主播黑名单、视频 MD5 库和直播间热度分析等来制定辅助检测功能； 1.1.3 能利用 OCR 技术（文字信息、行信息、行位置信息、背景检测等）、人脸识别技术（人脸个数、人脸姓名、人脸位置信息、AI 换脸等）、质量检测（图片大小、图片尺寸、美观度、边框检测）等加强图片检测辅助能力； 1.1.4 能运用 识别引擎、声纹识别引擎（娇喘检测、ASMR 检测）和语种检测引擎（英语、日语、韩语等）提升音频检测正确率。
	1.2 人工检测	1.2.1 能掌握人工审核与智能审核的差异性和互补性，进行专业审核团队的组建与管理，并建立应急响应机制； 1.2.2 能利用丰富的审核经验与场景，制定完善的舆情培训体系，及时调整落实最新监管政策； 1.2.3 能运用和搭建审核体系，灵活增加或削减审核人力，支持定制化人力调配需求。

2.智能审核管理系统	2.1 智能审核管理系统的优化	2.1.1 能依托影像系统、以及智能审核平台对特殊场景需求进行互联网内容审核解决方案和功能优化； 2.1.2 能正确的掌握审核规则，利用流程化、标准化、自动化等技术，进行审核成本和运维成本的的缩减； 2.1.3 能依据多部门、多平台发布公开内容的需求，严格规范信息发布流程，审核流程可溯源、审核结果可统计。
	2.2 审核结果分析	2.2.1 能利用大数据分析技术，实时掌握内容风险动态； 2.2.2 能掌握可视化结果反馈设计原理，针对智能审核结果，可从具体规则、时间、区域、模块、业务范围等多个维度去分析单据的审核情况，通过分析用户习惯以及规则驳回原因，以此提升单据填报的通过质量，也可以作为共享审核人员工作调配的依据； 2.2.3 能运用实时有害信息分布情况和内容安全指标的查看，进行安全风险的把控。
3. 互联网内容审核项目管理	3.1 传达审核组任务	3.1.1 能正确了解审核组的任务，及时发现组内存在的问题，能够有效沟通并提供解决方案： 3.1.2 能掌握内容审核任务要求，确定任务有效地建立、实施并保持了内容体系过程和程序，以便为建立对内容审核体系的信任提供基础； 3.1.3 能熟悉审核任务传达的流程，以便告知客户其方针、目标及指标(与相关内容体系标准或其他规范性文件的期望一致)与结果之间的任何不一致，以使其采取措施。

	3.2 项目管理	3.2.1 能正确了解审核组的任务，及时发现组内存在的问题，能够有效沟通并提供解决方案： 3.2.2 能掌握内容审核任务要求，确定任务有效地建立、实施并保持了内容体系过程和程序，以便为建立对内容审核体系的信任提供基础； 3.2.3 能熟悉审核任务传达的流程，以便告知客户其方针、目标及指标(与相关内容体系标准或其他规范性文件的期望一致)与结果之间的任何不一致，以使其采取措施。
	3.3 培训与指导	3.3.1 能掌握互联网内容审核的知识与技能，进行初级、中级人员理论知识和操作技能的培训； 3.3.2 能利用审核的案例与经验，编写互联网内容审核项目培训计划、技术文件和培训课件； 3.3.3 能运用不同特色应用场景，指导初级、中级人员进行审核错误原因分析和故障排除。
4.审核方案架构与策略配置	4.1 制定审核计划	4.1.1 能掌握审核策略与人员组织，制定审核方案，清晰地确定所需要的审核活动； 4.1.2 熟悉相关审核评估标准、适用的法律法规等，进行审核目标制定、审核规则配置、审核人员配置，并落实到智能审核管理平台、人机协同审核管理平台中； 4.1.3 能运用审核目标、审核范围和审核计划，进行内容审核，并实现其预定目标的有效性衡量；

	4.2 提交审核报告	4.2.1 能掌握审核报告的内容要求，确保审核报告得到编制，且对审核报告的内容负责； 4.2.2 能利用互联网内容审核报告，进行审核人员的培训与智能审核系统的升级改进； 4.2.3 能掌握审核范围、审核模式及所遇风险情况，定期制作互联网内容审核风控报告。

参考文献

[1]GB/T 19001-2016 质量管理体系要求

[2]GB/T 19000-2016 质量管理体系基础和术语

[3]中华人民共和国主席令第五十三号《中华人民共和国网络安全法》

[4]国信办通字〔2019〕3 号《网络音视频信息服务管理规定》

[5]工业和信息化《信息通信行业发展规划(2016-2020 年)》

[6]《“十三五”国家信息化规划》

[7]《国家信息化发展战略纲要》

[8]《网络短视频内容审核标准细则》

[9]《网络信息内容生态治理规定》

[10]《微博客信息服务管理规定》